DOCUMENT RESUME

ED 459 199 TM 033 496

AUTHOR	Milewski, Glenn B.; Patelis, Thanos
TITLE	Factorial Invariance of the Advanced Placement International English Language Exam (APIEL[R]) across Chinese and German Samples.
PUB DATE	2000-10-00
NOTE	26p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (31st, Ellenville, NY, October 25-27, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Advanced Placement; *English (Second Language); *Factor Structure; Foreign Countries; *High School Students; High Schools; *Language Proficiency
IDENTIFIERS	Advanced Placement Examinations (CEEB); China; *Chinese People; Confirmatory Factor Analysis; *Germans; Germany; International English; Invariance

ABSTRACT

This study used confirmatory factor analytic methods to investigate whether the subscales of the Advanced Placement International English Language examination (APIEL[R]) measuring Writing, Speaking, Listening, and Reading were invariant across 2 groups, 197 Chinese students and 434 German students. Since categorical responses were observed on APIEL[R] indicators, models were fit to the polychoric correlation matrix and the matrix of asymptotic variances and covariances. PRELIS 2 was used to estimate the correlation matrix and the asymptotic variances and covariances of the examined indicators; LISREL 8 was used to perform confirmatory factor analysis (K. Joreskog and D. Sorbom, 1996). Results indicate that while the factors comprising the APIEL[R] are valid across groups, examinees did not interpret the content of indicators equivalently across groups. The paper discusses limitations of this research and next steps to take. (Contains 2 figures, 4 tables, and 19 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running head: STRUCTURE OF AN ENGLISH LANGUAGE PROFICIENCY TEST

ED 459 199

Factorial Invariance of the Advanced Placement International English Language exam (APIEL®) across Chinese and German samples.

Glenn B. Milewski

Fordham University

Thanos Patelis

The College Board

Paper presented at the annual meeting of the Northeastern Educational Research Association at Ellenville, New York, October 2000.

TM033496

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
G. Milewski
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

BEST COPY AVAILABLE

Abstract

The current study utilized confirmatory factor analytic methods to investigate whether the sub-scales of the Advanced Placement International English Language exam (APIEL) measuring Writing, Speaking, Listening and Reading were invariant across two groups. Chinese (n=197) and German (n=434) students comprised the two groups sampled. Since categorical responses were observed on APIEL indicators, models were fit to the polychoric correlation matrix and the matrix of asymptotic variances and covariances. PRELIS 2 was used to estimate the correlation matrix and the asymptotic variances and covariances of the exam indicators; LISREL 8 was used to perform confirmatory factor analyses (Jöreskog & Sörbom, 1996). Results indicated that while the factors comprising the APIEL are valid across groups, examinees did not interpret the content of indicators equivalently across groups. Limitations and next steps are discussed.

The Factor Structure of an English Language Proficiency Test

This study was undertaken to provide some information about the structural equivalence on an English language proficiency test, the Advanced Placement International English Language (APIEL) test, across two cultures. As this test continues to grow in its use internationally, evidence about its validity with examinees of global representation is needed. Therefore, this study was undertaken to begin the empirical examination of the structure of the APIEL in examinees from Germany and China.

In the growing literature addressing the adaptation of tests into multiple languages and cultures, there has been an increasing concern about the appropriateness of tests used in multiple languages and cultures (van de Vijver & Poortinga, 1991). As a result of this growing concern, guidelines were recently prepared by an international committee of psychologists under the initiative of the International Test Commission (see, Hambleton, 1994; van de Vijver & Hambleton, 1996).

Common errors may occur in the test adaptation process. Hambleton (1999) outlined four aspects involving the entire assessment process that affect the valid inference of scores: (1) construct equivalence, (2) test administration, (3) test format, and (4) speededness. Hambleton described construct equivalence as the prerequisite for doing any cross-national, cross-cultural, or cross-language comparisons. Therefore, this study examined the construct equivalence of an English language proficiency examination.

Various methods have been used to evaluate construct equivalence. Hui and Triandis (1985) suggest regression methods, item response theory (IRT) approaches, factor analyses, and multidimensional scaling. Examples of applying factor analytic techniques to the examination of cross-cultural construct equivalence exist (see van de Vijner and Leung, 1997).

Of particular interest to the authors is the application of confirmatory factor analysis. Because the purpose of this research was to examine whether the intended factors were represented by the examinees' performance, confirmatory factor analyses (CFA) were considered most appropriate. Examples of the application of CFA in the examination of construct equivalence are found in the literature (e.g., Rock & Werts, 1979; Sireci, Fitzgerald, & Xing, 1998; Everson, Guerrero, & Laitusis, 1998).

The instrument used in the current study was the Advanced Placement International English Language (APIEL) exam. The APIEL exam measures four constructs related to English language proficiency: Listening, Reading, Writing, and Speaking. The exam was designed to identify non-native speakers who can use English well enough to participate in regular classes at an English-speaking university (The College Board, 1997). Listening and reading skills are measured by multiple-choice questions, while writing and speaking skills are assessed by free-response items.

The purpose of this study is to compare the factor structure of the APIEL exam across two groups of examinees. Using multi-group confirmatory factor analysis, this study will examine whether the factor structure is consistent between examinees from China and Germany. As the expansion of an English language proficiency test moves into other countries across the world, there is concern whether the constructs that were intended for this test are similar for groups of examinees from different parts of the world.

Several hypotheses were examined in the current study. First, the theoretical 4-factor model was examined in each ethnic group separately. Next, the invariance of the 4-factor solution across groups was investigated. Finally, the invariance of the pattern of factor loadings was examined. Since the English language test used in this study contains items that provide

categorical responses, all hypothesized models were fit to the polychoric correlation matrix of indicators and the asymptotic variances and covariances of those indicators.

Method

Participants

The present sample comprised 197 Chinese students and 434 German students. Exams were administered to students in their native countries. Chinese students were examined in May of 1999 whereas German students were examined in May of 1998. Examinees differed with respect to AP grades on the exam (Chinese students, $\underline{M} = 2.62$, $\underline{SD} = 1.04$; German students, $\underline{M} = 3.15$, $\underline{SD} = 1.20$) and scores on sub-scales (see Table 1).

Instrumentation and Procedure

The College Board first introduced the Advanced Placement International English Language (APIEL) examination in 1997. The APIEL is comprised of two sections: multiple choice and free response. The multiple choice section is made up of a Listening Comprehension sub-scale containing 41 items measured in 35 minutes and a Reading Comprehension sub-scale containing 39 items measured in 50 minutes. Each multiple choice item consists of an item stem and 4 choices. The free response section also contains 2 sub-scales, Writing, which is made up of two 40-minute essay questions and Speaking, which contains 5 questions measured in a total of 15 minutes. Speaking questions are scored on a 5-point scale. Essay question 1 is scored on a 10-point scale and essay question 2 is scored on a 15-point scale. The total test score for the APIEL examination is a weighted composite of scores on Sections I and II expressed on a 1 to 5 scale (Educational Testing Service, 2000).

The administration of these tests was conducted as part of the annual administration of AP examinations throughout the world. Examinees from both countries were administered the

same form of the test. However, the examinees from Germany took the test in 1998 and the examinees from China took the test in 1999. In 1998 and 1999, there were 3,752 and 4,633 total examinees throughout the world, respectively.

Likewise, the scoring was performed in the annual scoring sessions immediately following the administration. Thus, the tests for the examinees from Germany and China were scored in 1998 and 1999, respectively. Comparability of scores is accomplished by utilizing common-item equating of the multiple-choice items, and an annual AP grade setting.

Multiple-choice responses are scored electronically. The multiple-choice items are formula scored with a deduction of a fraction of a point for each incorrect response. Free-response questions are scored using human scorers shortly after the administration in June. The scorers represent school and University teachers from various institutions across the world. Scorers are selected based on school locale and setting, gender, ethnicity, and years of teaching experience. A Chief Faculty Consultant (CFC), appointed for a four-year term, after serving for one year as a CFC Designate, (1) supervises the scoring of the free-response section of the exam, (2) acting as a major contributor to the development of the examination, and (3) communicates to the scoring committee how the candidates responded to and performed on the free-response portions of the exams.

The scoring of the free-response involves an extensive process. During the creation of the free-response questions, preliminary scoring standards are produced. Before the actual scoring takes place, the CFC prepares a draft of the scoring guidelines for each free-response question. Next, immediately prior to scoring the CFC and various key test developers and scorers to review and revise the draft scoring guidelines, and test them by prescoring randomly selected student papers. Afterwards, the CFC and key scorers conduct training sessions for each free-

response question, which are attended by all scorers. A scoring reliability study conducted in 1998 found the scoring consistency[1] between the operational scoring and a second experimental scoring of a sample of 222 exams was 0.82.

The multiple-choice items showed an internal consistency of 0.89 using KR-20. For the composite score, the lower-bound reliability estimate was 0.88 and the upper-bound estimate was 0.91 (Educational Testing Service, 1999).

Confirmatory factor analysis models were fit to the polychoric correlation matrices for Chinese and German examinees (See Tables 2 and 3). PRELIS 2.3 was used to estimate the correlation matrix and the asymptotic variances and covariances of the exam indicators; LISREL 8.3 was used to perform confirmatory factor analyses (Jöreskog & Sörbom, 1996).

Results

Using confirmatory factor analytic (CFA) procedures (LISREL 8.3; Jöreskog & Sörbom, 1999), the data were analyzed in two stages. First, the factorial validity of the APIEL was tested separately for Chinese and German examinees. Model specifications and parameter estimates are provided (See Figures 1 & 2). Second, the factorial invariance was of the APIEL was tested across the two ethnic groups sampled. Analyses were conducted on both single items and parcels of items. In accordance with recommendations provided by MacCullum and Austin (2000) and Hu and Bentler (1998), assessment of model fit was based on the Standardized Root Mean Square Residual (SRMSR), the Root mean Square Error of Approximation (RMSEA), and the Non-Normed Fit Index (NNFI). Chi-square ($\underline{X}^2$) values were also reported but only used to evaluate the fit of nested models because of well known problems associated with the influence of sample size and other variables on chi-square values (Bentler & Bonnet, 1980; Marsh & Hocevar, 1985; Hu & Bentler, 1988).

[1] Reader reliability was calculated as total variance-error variance / total variance.

The CFA model in the present study hypothesized a priori that: (a) each indicator had a non-zero loading on the APIEL factor it was designed to measure, and zero loadings on all other factors, (b) responses to the APIEL exam were explained by 4 factors which all loaded on a single higher-order factor and (c) the APIEL exam was factorial invariant across groups.

The hypothesized 4-factor model represented a statistically acceptable fit to data derived from the Chinese sample. Various indices supported the tenability of the hypothesized model. For example, chi-square indicated a reasonable fit between the unrestricted sample polychoric correlation matrix and the restricted polychoric correlation matrix ($\underline{X}^2_{(61)} = 45.34$, $\underline{p} = .93$). The SRMSR was adequate (.79). Likewise, the RMSEA value for the hypothesized model was 0.0, with the 90 percent confidence interval ranging from 0.0 to 0.01 indicating that the model would provide a good fit to the population polychoric correlation matrix if it was available (Browne & Cudeck, 1993). Other fit indices also supported the 4 factor model (e.g. NNFI = 1.01). However, all fit indices are interpreted with caution due to the fact that several negative error variances, Heywood cases, were estimated when the single group model was fit to the sample of Chinese examinees.

The hypothesized 4-factor model did not provide an acceptable fit to the data from the German examinees ($\underline{X}^2_{(61)} = 203.54$, $\underline{p} < .05$; RMSEA = .073; 90% C.I. for RMSEA = .062 to .085). However, after freeing the error covariance between SPEAK4 and SPEAK5, the model did provide an acceptable fit. The RMSEA of the new model was (.055) and the 90% confidence interval for this index (.044 to .067) fell within the normal range (Byrne, 1998). Likewise other indices supported the fit of the model (e.g. NNFI = .99; SRMSR = .071).

Using the methodology outlined by Byrne (1998), multi-group invariance of the APIEL was investigated by testing a series of increasingly more restrictive hypotheses (see Table 4).

Hypothesis 1, which tested the validity of the 4-factor structure, was supported. Most results indicated a relatively good fit of the model to the data (SRMSR = .038; RMSEA = .041; 90% C.I. for RMSEA = .029 to .053; NNFI = .99). However, chi-square was significant ($X^2_{(121)}$ = 185.22, $p < .05$). Hypothesis 2, which tested the invariance of the factor loadings, was not supported ($\Delta X^2_{(9)} = 135.98$, $p < .05$)[2]. Based on the outcome of hypothesis 2, the invariance of each factor loading was investigated individually. It was determined, based on evaluations of the change in chi-square, that 3 of the 13 factor loadings were invariant. The following factor loadings were invariant across groups: (a) the loading of Speaking on SPEAK2 ($\Delta X^2_{(1)} = 2.16$, $p > .05$), (b) the loading of Listening on ListeningPARCEL3 ($\Delta X^2_{(1)} = 0.9$, $p > .05$) and (c) the loading of Reading on ReadingPARCEL3 ($\Delta X^2_{(1)} = 1.19$, $p > .05$).

Discussion

LISREL CFA procedures were used to test the factorial validity of the sub-scales of the APIEL exam. The results demonstrated a well-defined factor structure yielding one general APIEL factor, and 4 domains measuring English language proficiency- Writing, Speaking, Listening and Reading. This factor structure was invariant across groups as revealed by fit indices for Model 1. However the ΔX^2, representing the difference between Models 1 and 2 ($\Delta X^2_{(9)} = 135.98$), indicated that the pattern of factor loadings was not invariant across groups. Since the equality of the factor loading matrix was not tenable, it was necessary to test for the invariance of each of its individual parameters (Byrne, 1998). Tests indicated that only 3 of the 13 factor loadings were invariant across groups.

The major finding was that while the APIEL measured the same sub-scales, it did not measure these constructs in the same units across groups (i.e. non-invariant factor loadings).

---

[2] Change in chi-square (ΔX^2) was the major criteria used to evaluate the test of invariant factor loadings because the

Only 3 of the 13 indicators measured English proficiency in exactly the same way for both Chinese and German examinees. Examinees responded to only one indicator for each of the Speaking, Listening and Reading sections in the same way across groups; participants did not interpret either of the indicators for the Writing section equivalently. This finding implies that the strength with which most of the indicators measure the latent traits is different across the two groups sampled. In general, the factor loadings of the latent traits on the indicators were higher in the sample of Chinese examinees. The differential factor loadings across groups suggest that the APIEL measured English language proficiency more accurately in the sample of students from China.

Cultural differences in English teaching styles between China and Germany, (e.g., fragmented v. whole-language, respectively) may help to explain the greater degree of measurement accuracy in the sample of examinees from China. Other possible explanations of the higher factor loadings observed in the sample of students from China include: (1) the high degree of dissimilarity between the Chinese and English language and (2) the greater exposure to English in Germany (e.g. bilingual schools, proximity to English-speaking countries, introduction of English at 5th grade, etc.). These possibilities may explain why indicators measured latent traits more precisely in the sample of examinees from China.

The results of the current study are limited for several reasons. First, since the sampled groups differed in the scores achieved on the test, invariance was evaluated across groups that differed in both culture and ability. It is therefore more difficult to find invariance in this situation. It may be worthwhile to re-test these hypotheses controlling for differences in ability across groups. Second, the validity of our conclusions is threatened since examinees were tested 1 year apart. Third, while Heywood cases were not estimated in either of the multi-group CFA

---

model used to test this hypothesis was nested within the model used to test the invariance of the 4-factor solution.

models, they were estimated in the single group model fit to the Chinese sample. Heywood cases pose serious problems regarding the accuracy of parameter estimates and fit indices. As a consequence the results of the single group CFA on the Chinese sample are interpreted with caution. Further work is necessary in order to determine the effect, in any, that these estimates had on the results of the multi-group analyses.

Based on the results of this study, the next steps would be first to expand the comparison to examinees from other languages. For example, in order to test the hypothesis that the high degree of language dissimilarity may have affected the factor loadings, the addition of students from France (i.e., closer to students from Germany) and Japan (i.e., closer to students from China) may provide insight.

A final suggestion would be to identify the instructional technique utilized by the teachers of these students, and utilize this information in the analysis of the structural equivalence across cultures and language groups. This may provide evidence about the instructional effect on the observed factor structure.

References

Bentler, P. M. & Bonnet, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 445-455).

Byrne, B. (1998). Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Educational Testing Service. (1999). College Board Advance Placement International English Language Form 3UBP Reader Reliability Study. Unpublished report.

Educational Testing Service. (2000). 1999 Advanced placement test analyses: Forms 3vbp. (ETS Rep. No. SR-2000-07). Princeton, NJ: Educational Testing Service

Everson, H. T., Guerrero, A., & Laitusis, V. (1998, April). Preliminary evidence of construct equivalence of mathematics tests administered across languages: An analysis of findings from the SAT I and the Prueba de Aptitud Academica tests. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Hambleton, R. K. (1999). Issues, designs, and technical guidelines for adapting tests in multiple languages and cultures. Laboratory of Psychometric and Evaluative Research Report No. 353. Amherst, MA: University of Massachusetts, School of Education.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. European Journal of Psychological Assessment, 10, 229-240.

Hu, L. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, 3, 424-453.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. Journal of Cross-Cultural Psychology, 16, 131-152.

Jöreskog, K. & Sörbom, D. (1996). LISREL 8: User reference guide. Chicago, IL: Scientific Software International, Inc.

MacCullum, R. C. & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. Annual Review of Psychology, 53, 201-219.

Marsh, H. W., Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First and higher order factor models and their invariance across groups. Psychological Bulletin, 97, 562-582.

Rock, D. A., & Werts, C. E. (April, 1979). Construct validity of the SAT across populations: An empirical confirmatory study (RR-79-2). Princeton, NJ: Educational Testing Service.

Sireci, S. G., Fitzgerald, C., & Xing, D. (1998, April). Adapting credentialing examinations for international uses. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

The College Board. (1997). Advanced Placement International English Language. New York, NY: The College Board.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. European Psychologist, 1, 89-99.

van de Vijner, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandley (Eds.), Handbook of cross-cultural psychology (2nd ed., vol. 1, pp.257-300). Boston: Allyn & Bacon.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R.K. Hambleton & J. Zaal (Eds.), Advances in educational and psychological testing. Dordrecht: Kluwer Academic Publishers.

Author's Note

Glenn B. Milewski, Department of Psychology, Fordham University.

Thanos Patelis, Research and Development, The College Board.

A number of valuable and helpful discussions on technical issues related to structural equation modeling with Maureen Ewing, as well as the assistance of ETS staff members Rick Morgan and Behroz Maneckshana, are gratefully acknowledged. The authors would like to also thank Professor Yang Huizhong for his efforts organizing data collection for the Chinese examinees.

Correspondence regarding this paper should be addressed to the first author and sent to Fordham University, Dealy Hall, Department of Psychology, 441 East Fordham Road, Bronx NY, 10458. E-mail correspondence can be addressed to milewski@fordham.edu.

Table 1

Descriptive Statistics for APIEL Sub-scales

	Chinese Examinees				German Examinees			
Sub-scale	N	M	SD	Kurtosis	N	M	SD	Kurtosis
ESSAY1	197	3.45	0.89	-0.27	434	3.73	1.16	-0.39
ESSAY2	197	4.50	2.88	-1.22	434	4.67	1.82	-0.48
SPEAK1	197	2.73	1.05	-0.17	434	3.00	0.99	0.24
SPEAK2	197	2.82	0.92	0.22	434	3.30	0.91	0.99
SPEAK3	197	2.84	1.00	0.08	434	3.10	0.93	0.50
SPEAK4	197	2.63	1.05	-0.46	434	2.25	1.13	-0.16
SPEAK5	197	2.85	1.08	-0.28	434	3.60	1.06	0.21
Listening PARCEL1	197	10.09	2.19	1.15	434	10.69	2.24	-0.11
Listening PARCEL2	197	9.36	2.81	-0.33	434	9.84	2.32	0.76
Listening PARCEL3	197	8.83	2.37	-0.34	434	9.51	2.34	-0.47
Reading PARCEL1	197	9.45	2.05	3.65	434	9.55	2.13	0.48
Reading PARCEL2	197	7.53	2.39	-0.03	434	8.07	2.76	-0.17
Reading PARCEL3	197	9.37	2.05	3.22	434	9.77	2.22	0.31

Table 2
Polychoric Correlation Matrix (Chinese Examinees)

	Essay1	Essay2	Speak1	Speak2	Speak3	Speak4	Speak5	Listen1	Listen2	Listen3	Read1	Read2	Read3
Essay1	1.00												
Essay2	-0.10	1.00											
Speak1	0.44	-0.05	1.00										
Speak2	0.40	-0.10	0.76	1.00									
Speak3	0.32	-0.06	0.44	0.39	1.00								
Speak4	0.45	-0.10	0.30	0.41	0.60	1.00							
Speak5	0.40	0.00	0.32	0.25	0.57	0.55	1.00						
Listening1	0.46	0.03	0.23	0.16	0.39	0.45	0.44	1.00					
Listening2	0.38	0.14	0.28	0.19	0.37	0.34	0.35	0.50	1.00				
Listening3	0.22	0.19	0.16	0.03	0.30	0.39	0.41	0.51	0.45	1.00			
Reading1	0.13	0.03	0.38	0.34	0.19	0.42	0.10	0.00	0.09	0.05	1.00		
Reading2	-0.06	0.86	-0.08	-0.08	0.06	-0.01	0.07	0.06	0.20	0.20	0.09	1.00	
Reading3	0.15	-0.13	0.20	0.20	0.17	0.31	0.04	-0.05	0.08	0.08	0.48	-0.10	1.00

Table 3
Polychoric Correlation Matrix (German Examinees)

	Essay1	Essay2	Speak1	Speak2	Speak3	Speak4	Speak5	Listen1	Listen2	Listen3	Read1	Read2	Read3
Essay1	1.00												
Essay2	0.45	1.00											
Speak1	0.26	0.26	1.00										
Speak2	0.30	0.27	0.74	1.00									
Speak3	0.29	0.28	0.74	0.84	1.00								
Speak4	0.35	0.23	0.49	0.50	0.57	1.00							
Speak5	0.40	0.32	0.52	0.57	0.60	0.80	1.00						
Listening1	0.32	0.35	0.28	0.40	0.36	0.38	0.45	1.00					
Listening2	0.28	0.31	0.29	0.38	0.35	0.32	0.42	0.62	1.00				
Listening3	0.30	0.25	0.28	0.36	0.36	0.32	0.39	0.63	0.65	1.00			
Reading1	0.34	0.31	0.23	0.31	0.32	0.26	0.33	0.53	0.49	0.53	1.00		
Reading2	0.27	0.32	0.15	0.22	0.25	0.16	0.32	0.47	0.50	0.50	0.59	1.00	
Reading3	0.31	0.30	0.17	0.25	0.23	0.24	0.32	0.50	0.46	0.51	0.66	0.64	1.00

Table 4

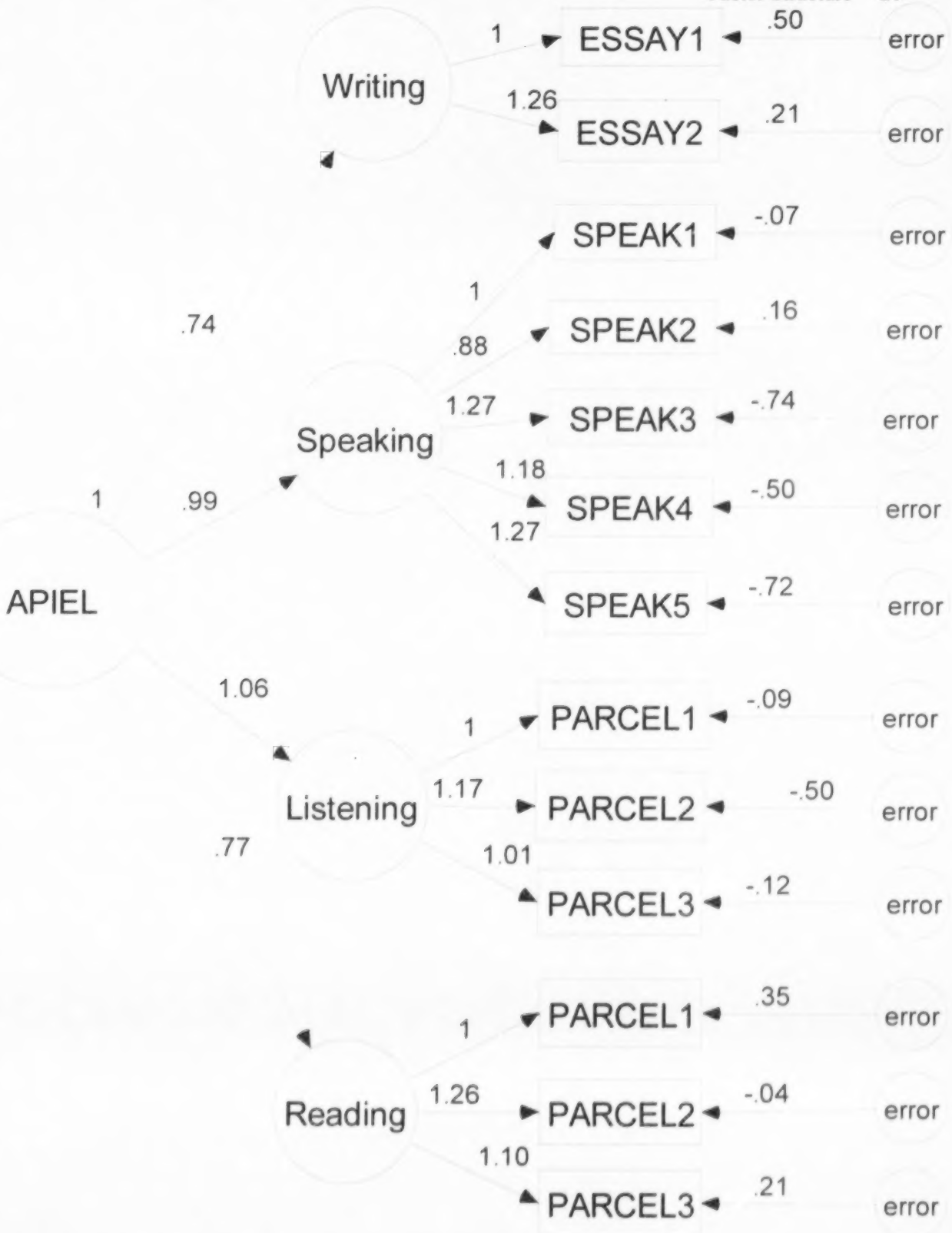
Pertinent Results for Tests of Multi-Group Factorial Invariance of the APIEL Exam

Hypothesis	X^2	(df)	ΔX^2	(df)	SRMSR	RMSEA	NNFI
1. Number of factors invariant	185.22	(121)	----	----	.038	.041	0.99
2. Model 1 with pattern of factor loadings held invariant	321.20	(130)	135.98	(9)	.054	.068	0.98

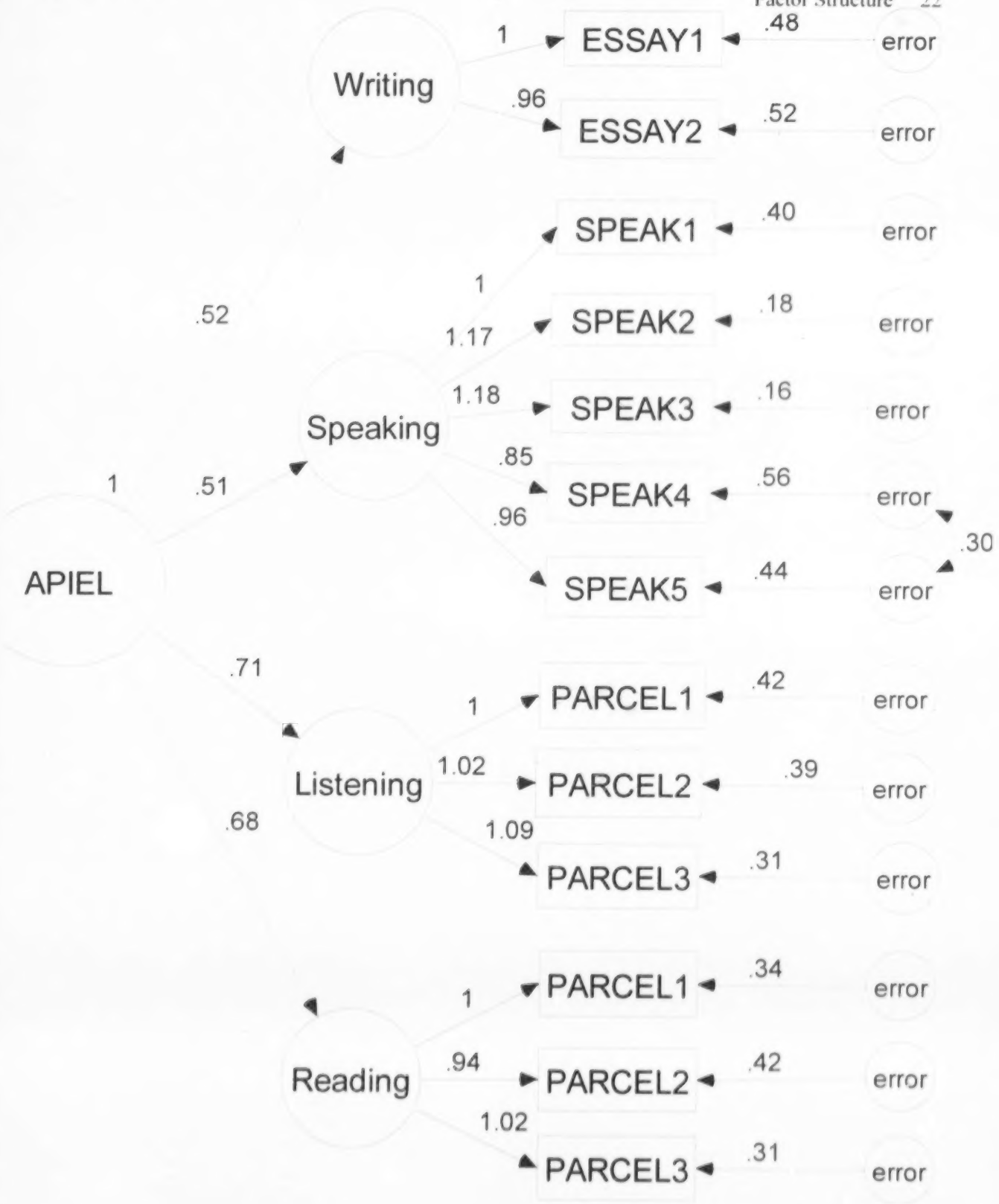
Figure Caption

Figure 1. Path diagram and parameter estimates for the hypothesized 4-factor model for Chinese Examinees.

Figure 2. Path diagram and parameter estimates for the hypothesized 4-factor model for German Examinees.



BEST COPY AVAILABLE



BEST COPY AVAILABLE